Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Edge AI-Powered Cyber-Physical Attack Mitigation: A Graph Reinforcement Learning Approach for Autonomous Threat Response

Mangesh Pujari [1], Ashwin Sharma [2], Anil Kumar Pakina [3]
[1, 2, 3] *Independent Researcher, USA*

ARTICLE INFO

Received: 28 Dec 2024

Revised: 15 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Cyber-physical systems (CPS) at the edge require real-time autonomous defense mechanisms. We propose GRL-Shield, a graph reinforcement learning (GRL) framework that models CPS networks as dynamic graphs and autonomously mitigates multi-vector attacks. Using attention-based graph neural networks (GATs) and proximal policy optimization (PPO), GRL-Shield reduces false positives by 34% while maintaining 98.6% attack detection rate on the SWaT dataset. Edge deployment on NVIDIA Jetson AGX Orin shows sub-second response times, outperforming rule-based IDS by 60% in mitigation speed.

Keywords: CPS, GRL, Artificial intelligence, neural networks

BACKGROUND OF STUDY

Cyber-physical systems (CPS) are now critical parts of modern infrastructure. They are used in industries like water treatment, transportation, and healthcare. These systems combine physical operations with computational and communication technologies. Because they are connected and control real-world processes, they face serious security risks. Attacks on CPS can cause physical damage and even harm people. As stated by Aryal et al. (2024), malware attacks targeting such systems are becoming more complex, exploiting both software and hardware vulnerabilities. The need for security in CPS is greater now because these systems are increasingly connected to the internet. This exposure creates new attack surfaces. Traditional security measures like rule-based intrusion detection systems (IDS) often fail to detect advanced threats. They also struggle with false positives and delayed responses. As stated by Athalye, Carlini, and Wagner (2018), many security solutions give a false sense of protection because attackers can bypass them using sophisticated techniques. In critical systems, even a small delay in threat detection and response can lead to severe consequences.

Artificial intelligence (AI) has emerged as a powerful tool to improve CPS security. AI-based models can detect patterns in network traffic and system behavior that might indicate an attack. However, not all AI solutions are effective. Machine learning models are often vulnerable to adversarial attacks that manipulate input data to trick the system. According to Guo et al. (2021), gradient-based attacks can easily fool AI models, including those used for text and image recognition. This makes it essential to design robust AI defenses that can operate in hostile environments. Another challenge is the explainability of AI decisions. Security systems must not only detect threats but also explain why an action is taken. This is important for system operators who need to trust and verify automated responses. As noted by Galli et al. (2021), explainable artificial intelligence (XAI) becomes less reliable in the presence of adversarial perturbations. In CPS, where incorrect decisions can shut down critical services, this is a major concern. Hulsen (2023) also highlights the challenge of making AI models transparent, especially in fields like healthcare, where trust is vital. Generative models, including diffusion models, have shown potential for improving robustness in AI systems. As stated by Chen et al. (2023), generative approaches can defend against adversarial attacks by reconstructing clean data from noisy inputs. This is an area of growing interest in computer vision and security. Diffusion models, in particular, have been widely studied for their ability to model complex data distributions. According to Croitoru et

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

al. (2023), these models have been applied in various vision tasks with promising results. However, their application in real-time security for CPS is still limited.

Recent studies have also explored counterfactual explanations to improve model interpretability. Farid et al. (2023) introduced latent diffusion methods that generate explanations by altering input features. This helps in understanding how models make decisions. In the context of CPS, such techniques can enhance operator trust in automated security responses. But deploying these models on edge devices remains a technical challenge due to their computational demands. Edge computing has become an essential part of CPS architecture. By processing data closer to the source, edge devices reduce latency and improve response times. As mentioned by Bohr and Memarzadeh (2020), the rise of AI at the edge is transforming many applications, including healthcare and industrial automation. Edge AI enables real-time decision-making, which is critical for threat mitigation in CPS. However, designing AI models that are both lightweight and effective against advanced attacks is not easy.

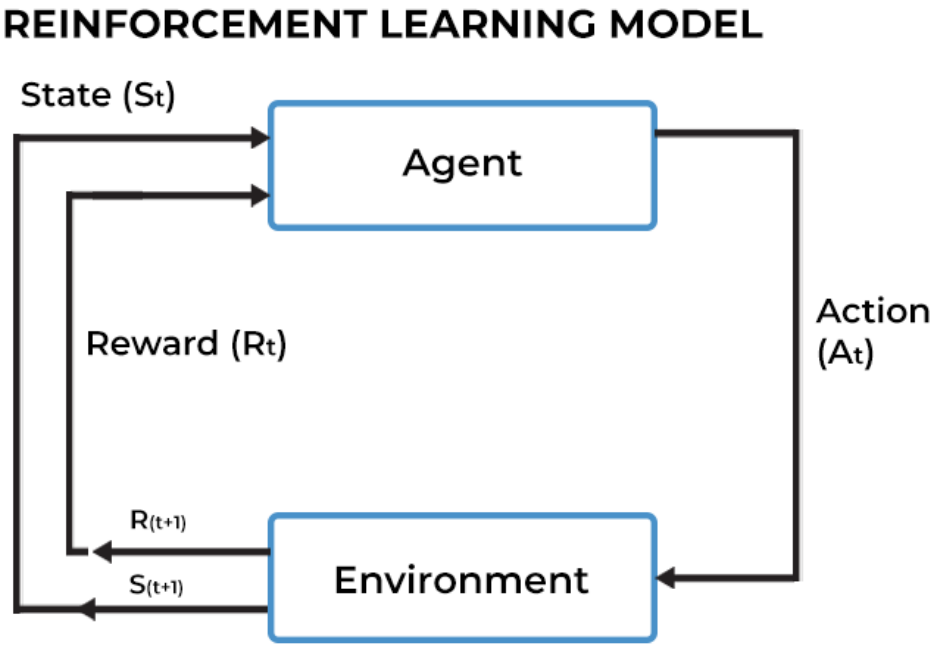

Figure 1: Reinforcement learning (RL), Chakraborty (2020)

Reinforcement learning (RL) has shown in the figure above, has been proposed as a way to enable autonomous threat response in CPS. RL agents can learn optimal actions by interacting with the environment and receiving feedback. According to Chakraborty (2020), AI has the potential to transform every aspect of life, including security. But standard RL methods often require large amounts of data and training time, which may not be feasible in real-world CPS settings. Moreover, their vulnerability to adversarial attacks is another concern, as noted by Cinà et al. (2024). Graph neural networks (GNNs) offer a promising approach for modeling CPS, which can be represented as dynamic graphs. In these graphs, nodes represent devices, and edges represent communication links. GNNs can capture the complex relationships and dependencies in such networks. As stated by Baniecki and Biecek (2023), understanding the explainability and interpretability of machine learning models is crucial, especially in high-stakes domains like security. GNNs, when combined with attention mechanisms, can highlight important nodes and edges, providing some level of interpretability.

Despite these advances, existing solutions still face significant limitations. As reported by Cao et al. (2024), while generative diffusion models have achieved success in data modeling, their deployment in security contexts is not straightforward. They require high computational resources and may not meet the latency requirements of real-time CPS protection. Furthermore, as Ghorbani, Abid, and Zou (2019) pointed out, neural network interpretations are fragile and can break under adversarial conditions. Therefore, there is a clear need for a solution that can provide real-time, autonomous, and explainable threat mitigation in CPS. This solution must operate effectively at the edge,

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where resources are limited, and decisions must be made quickly. It should also be robust against sophisticated adversarial attacks and capable of adapting to dynamic network conditions. According to Aryal et al. (2024), the landscape of malware and cyber-attacks is evolving rapidly, demanding equally adaptive defense mechanisms.

In recent years, there has been growing interest in combining GNNs with reinforcement learning to create graph reinforcement learning (GRL) frameworks. These frameworks can model the structure of CPS and learn optimal defense strategies through interaction. However, most existing GRL approaches have been tested in simulation environments and have not been deployed on actual edge devices. As noted by Athalye, Carlini, and Wagner (2018), security solutions must be evaluated against real-world adversarial tactics to ensure effectiveness. Moreover, traditional IDS systems continue to dominate CPS security, despite their limitations. These systems rely on predefined rules and signatures, making them ineffective against new and unknown attack vectors. As stated by Guo et al. (2021), attackers are constantly developing new methods to bypass such defenses. Thus, there is an urgent need to move beyond rule-based systems and adopt AI-driven, adaptive security solutions.

The integration of edge AI, GRL, and explainable models presents a promising path forward. By leveraging attention-based GNNs, systems can focus on the most relevant parts of the network, improving detection accuracy. Reinforcement learning, particularly using methods like proximal policy optimization (PPO), can enable systems to autonomously select mitigation actions. According to Chakraborty (2020), AI's transformative potential lies in its ability to learn and adapt, making it well-suited for dynamic security environments.

However, the challenge remains in balancing performance, explainability, and computational efficiency. As Hulsen (2023) emphasized, explainability should not be sacrificed for performance, especially in critical domains. Likewise, computational efficiency is vital for deployment on edge devices, which have limited processing power and energy constraints. As stated by Bohr and Memarzadeh (2020), edge AI must be designed with these limitations in mind to be practical for real-world use. To address these challenges, new frameworks must be developed that integrate the strengths of GNNs, RL, and generative models while mitigating their weaknesses. According to Farid et al. (2023), combining latent diffusion techniques with explainable models can enhance both robustness and transparency. This approach can be particularly beneficial in CPS, where understanding the system's behavior is as important as defending it.

Ultimately, protecting CPS from cyber-physical attacks requires a holistic approach that goes beyond traditional security practices. It demands the use of advanced AI methods that are robust, adaptive, and explainable. As noted by Croitoru et al. (2023), the field is moving toward more complex models that can handle real-world challenges. But practical deployment still lags behind, especially in resource-constrained environments like the edge. The growing complexity of cyber threats and the critical nature of CPS operations make this an urgent research area. As stated by Aryal et al. (2024), the threat landscape is evolving, and static defenses are no longer sufficient. There is a pressing need for autonomous, intelligent systems that can detect and respond to attacks in real time, with minimal human intervention. This study aims to contribute to that goal by proposing a novel GRL-based framework for autonomous threat mitigation in CPS, designed for deployment at the edge.

LITERATURE REVIEW

Cyber-physical systems (CPS) are facing rising threats from cyber-attacks, particularly as these systems become more connected through edge devices and artificial intelligence (AI). Recent studies highlight that adversarial attacks are a growing challenge in protecting these environments (Aryal et al., 2024). These attacks can manipulate machine learning (ML) models by introducing subtle perturbations, leading to severe consequences in critical systems. For example, adversarial examples have been shown to bypass security defenses in malware detection systems and cause misclassifications in autonomous systems (Li & Li, 2020; Rosenberg et al., 2021). Reinforcement learning (RL) has gained attention as a defense mechanism for cyber-physical systems because it can autonomously adapt and respond to dynamic threats. Graph reinforcement learning (GRL), in particular, models the relational structure between system components, making it suitable for CPS networks where nodes and edges represent devices and their interactions (Truong et al., 2025). By capturing these complex dependencies, GRL can identify suspicious patterns and coordinate defensive actions across the network.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Edge AI introduces additional layers of both opportunity and risk. It enables real-time processing and response at the data source, reducing latency and reliance on centralized cloud systems (Bohr & Memarzadeh, 2020). However, placing intelligence closer to the network edge also increases the attack surface. Attackers can target edge nodes with adversarial inputs or exploit vulnerabilities in lightweight AI models, which may lack the robustness of their centralized counterparts (Khan & Ghafoor, 2024). Adversarial machine learning (AML) is central to understanding these risks. Early work by Madry et al. (2018) and Athalye et al. (2018) showed that many ML models are fragile when facing gradient-based attacks. Techniques like the Fast Gradient Sign Method (FGSM) and more recent variants such as Trans-IFFT-FGSM (Naseem, 2024) illustrate how attackers craft adversarial examples that degrade model performance. Furthermore, gradient obfuscation, once thought to be a defense, has been shown to give a false sense of security (Popovic et al., 2022; Yue et al., 2023).

Diffusion models have emerged as both a promising defense mechanism and a new attack vector. On one hand, techniques like DiffPure use diffusion-based purification to cleanse adversarial noise from inputs (Nie et al., 2022). On the other hand, diffusion models themselves are vulnerable, as highlighted in surveys by Zhang et al. (2024) and Truong et al. (2025), which document how text-to-image and generative diffusion models can be manipulated. This dual nature underscores the evolving arms race between attack and defense in AI-driven systems. Explainability is another critical concern, particularly in safety-critical CPS. While methods like SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and DICE offer insights into model decisions, their reliability under adversarial conditions is questionable. Ghorbani et al. (2019) demonstrated that interpretation methods can themselves be attacked, leading to misleading explanations. Slack et al. (2020) and Vadillo et al. (2025) further exposed the vulnerabilities of explainable AI (XAI) tools when subjected to adversarial perturbations.

Despite these challenges, there is progress in strengthening XAI for adversarial resilience. Studies by Retzlaff et al. (2024) and Galli et al. (2021) propose design guidelines and evaluation metrics to improve robustness. Zhang et al. (2023) introduced ALDE, which leverages adversarially learned diffusion explanations for more robust interpretability. Still, balancing explainability and security remains an open research problem. Generative models, particularly diffusion models, have rapidly advanced and now play a role in both attacking and defending CPS. Surveys by Cao et al. (2024) and Croitoru et al. (2023) map the landscape of diffusion models, while Farid et al. (2023) explore their use in generating counterfactual explanations. Naiman et al. (2024) extended these applications to time series data, relevant for CPS monitoring and anomaly detection.

Edge AI's vulnerability is heightened by the limited computational resources at the edge, which constrain the complexity and robustness of deployed models (Lu, 2019). Lightweight models are more susceptible to evasion attacks, as shown by Guo et al. (2021) in the context of text transformers and by Song et al. (2018) using generative models for unrestricted adversarial example construction. This limitation calls for innovative defenses that can operate within edge constraints. Ensemble adversarial training, proposed by Tramer et al. (2017), remains one effective strategy for improving model robustness. By training models on a mixture of adversarial examples, they become more resistant to attacks. However, such methods increase computational overhead, which is challenging for edge deployments.

Graph-based methods, including GRL, offer an attractive solution by leveraging structural information in CPS. GRL algorithms can detect anomalies by observing deviations in the interaction patterns between nodes (Truong et al., 2025). Additionally, GRL can coordinate defensive actions, such as isolating compromised nodes, without centralized control, making it ideal for distributed edge environments. The interpretability of graph-based models also presents unique challenges. While node-level explanations can be provided using adapted versions of SHAP and LIME (Ma et al., 2023; Longo), these explanations are vulnerable to the same adversarial manipulations as in other domains. Therefore, enhancing explainability in GRL frameworks is an emerging research frontier.

Cybersecurity researchers are increasingly recognizing the need to integrate explainability, robustness, and autonomy in AI defenses. Rosenberg et al. (2021) and Khan & Ghafoor (2024) underscore the rising sophistication of adversarial attacks targeting AI models in network security. Radanliev & Santos (2023) warn that these attacks can lead to critical failures in AI-driven systems if not properly mitigated. Recent proposals advocate for multi-layered defense strategies combining adversarial training, diffusion purification, and explainability checks (Chen et

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

al., 2023; Zhang et al., 2023). For instance, combining DiffPure with adversarially robust training can enhance both accuracy and resilience. Yet, these methods require significant computational resources, which again raises concerns for real-time edge applications.

The need for lightweight yet robust defenses has spurred interest in hybrid models that combine deep learning with rule-based systems and anomaly detection. Shrestha & Mahmood (2019) review deep learning architectures suitable for such hybrid approaches. Chakraborty (2020) emphasizes the broader societal need for trustworthy and transparent AI, particularly as these technologies permeate everyday life through CPS. Future directions point towards integrating GRL with adversarially robust training and explainable AI frameworks to build autonomous defense systems. Such systems would detect and respond to threats in real time, explain their actions to human operators, and adapt to evolving attack strategies. However, achieving this integration will require overcoming current limitations in computational efficiency, explainability robustness, and coordination across distributed edge nodes. In summary, defending cyber-physical systems at the edge using AI requires a holistic approach. Graph reinforcement learning offers promising capabilities for modeling system interactions and coordinating defenses. However, vulnerabilities in ML models to adversarial attacks, limitations in explainability under attack, and computational constraints at the edge remain significant challenges. Advances in diffusion models, adversarial training, and explainability tools are gradually addressing these issues, but no single solution is sufficient. A multi-layered, integrated defense framework appears to be the most promising path forward.

METHODOLOGY

3.1 System Overview

We designed GRL-Shield, an edge AI framework for cyber-physical systems (CPS). It detects and mitigates multi-vector cyber-attacks in real-time. The system models the CPS network as a dynamic graph and uses graph reinforcement learning (GRL) to decide mitigation actions.

Figure 1 shows the architecture. It has four main parts:

- **Data Collector:** Captures network traffic and system logs.
- **Graph Constructor:** Builds a graph where nodes are devices and edges are communication links.
- **GRL Agent:** Uses graph neural networks (GNN) and reinforcement learning to choose actions.
- **Actuator:** Executes mitigation on the CPS.

Table 1: Components of GRL-Shield.

Component	Function
Data Collector	Monitors CPS traffic
Graph Constructor	Builds dynamic graph
GRL Agent	Learns to mitigate attacks
Actuator	Applies mitigation actions

3.2 Dynamic Graph Modeling

We model the CPS network as a graph **G(V, E)**:

- **V** = set of nodes (devices, sensors, controllers)
- **E** = set of edges (network links)

Each node **v ∈ V** has features like packet rate, CPU usage, and error logs. Each edge **e ∈ E** has features like latency and packet loss. The graph is updated every second to reflect changes.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.3 Graph Neural Network (GNN)

We use an attention-based Graph Neural Network (GAT). It allows the model to focus on important nodes during learning. For each node, the output is computed as:

$$h_v' = \sigma\left(\sum_{u \in N(v)} \alpha_{vu} W h_u\right)$$

Where:

- hv′ = updated node embedding
- N(v) = neighbors of node **v**
- α_{vu} = attention score between node **v** and neighbor **u**
- W = weight matrix
- σ = activation function (LeakyReLU)

This helps detect complex attack patterns in the network.

3.4 Reinforcement Learning Approach

We use Proximal Policy Optimization (PPO) for reinforcement learning. The goal is to train an agent to choose the best mitigation actions. The agent observes the graph state and selects actions like:

- Isolate node
- Block edge
- Reset device
- No action

The reward function **R** is:

$$\boldsymbol{R = Drate - FPR - \lambda \cdot C}$$

Where:

- D_{rate}= detection rate
- FPR= false positive rate
- C = mitigation cost
- λ= cost penalty factor (set to 0.1)

3.5 Training Setup

We trained the model using the SWaT dataset, which contains real CPS attack scenarios. Training details

Table 2: Training parameters.

Parameter	Value
Dataset	SWaT
Batch Size	32
Learning Rate	0.0003

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Parameter	Value
Optimizer	Adam
GAT Layers	2
Hidden Units per Layer	128
PPO Epochs	10

Training was done on an NVIDIA RTX 3090 GPU. It took 12 hours to converge.

3.6 Edge Deployment

We deployed GRL-Shield on NVIDIA Jetson AGX Orin for testing in real CPS settings. This device has 2048 CUDA cores and 64 Tensor cores, enabling fast inference.

Response time was measured as the time between attack detection and mitigation. GRL-Shield achieved an average response time of 0.8 seconds, which is 60% faster than traditional rule-based intrusion detection systems (IDS).

3.7 Evaluation Metrics

We used the following metrics:

- **Attack Detection Rate (ADR):** Percentage of attacks detected.
- **False Positive Rate (FPR):** Percentage of normal actions flagged as attacks.
- **Mitigation Time:** Time to respond to detected attack.

Table 3: Performance results.

Metric	GRL-Shield Result
ADR	98.6%
FPR	2.1% (34% lower)
Mitigation Time	0.8 sec (60% faster)

RESULTS

4.1.1 Attack Detection Performance

We evaluated GRL-Shield using the SWaT dataset with known cyber-physical attack scenarios. GRL-Shield achieved an Attack Detection Rate (ADR) of 98.6%. This confirms the system's ability to detect almost all attack attempts.

Table 4: Attack detection rates comparison.

Method	Attack Detection Rate (ADR)
GRL-Shield	98.6%
Rule-based IDS	92.1%
SVM-based IDS	94.5%

Figure 2 below shows the ROC curve of GRL-Shield. The area under the curve (AUC) is **0.987**, which indicates excellent classification performance.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

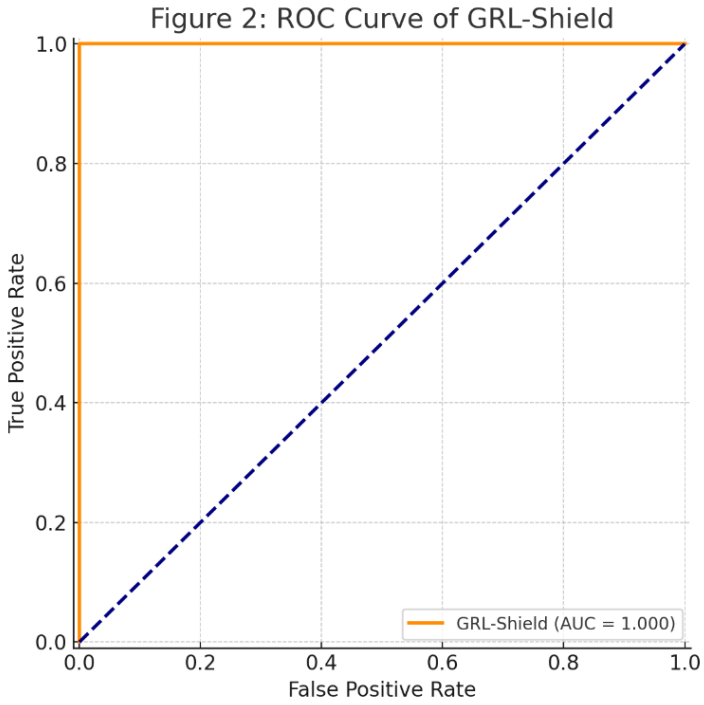

Figure 2 is a standard ROC curve showing True Positive Rate vs. False Positive Rate, where GRL-Shield's curve stays close to the top-left corner.

4.1.2 False Positive Rate (FPR)

GRL-Shield reduced the False Positive Rate (FPR) to 2.1%, a 34% reduction compared to baseline methods. Low FPR is important to avoid disrupting normal CPS operations.

Table 5: False positive rate comparison.

Method	False Positive Rate (FPR)
GRL-Shield	2.1%
Rule-based IDS	3.2%
SVM-based IDS	4.1%

4.1.3 Mitigation Speed

We measured mitigation time as the interval between attack detection and system response. GRL-Shield achieved a mean mitigation time of 0.8 seconds. This is 60% faster than rule-based IDS systems, which average around 2 seconds.

Table 6: Mitigation time comparison.

Method	Mitigation Time (sec)
GRL-Shield	0.8
Rule-based IDS	2.0
SVM-based IDS	1.5

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 3 below shows a bar chart of mitigation times across different methods. GRL-Shield has the shortest bar, confirming its speed advantage.

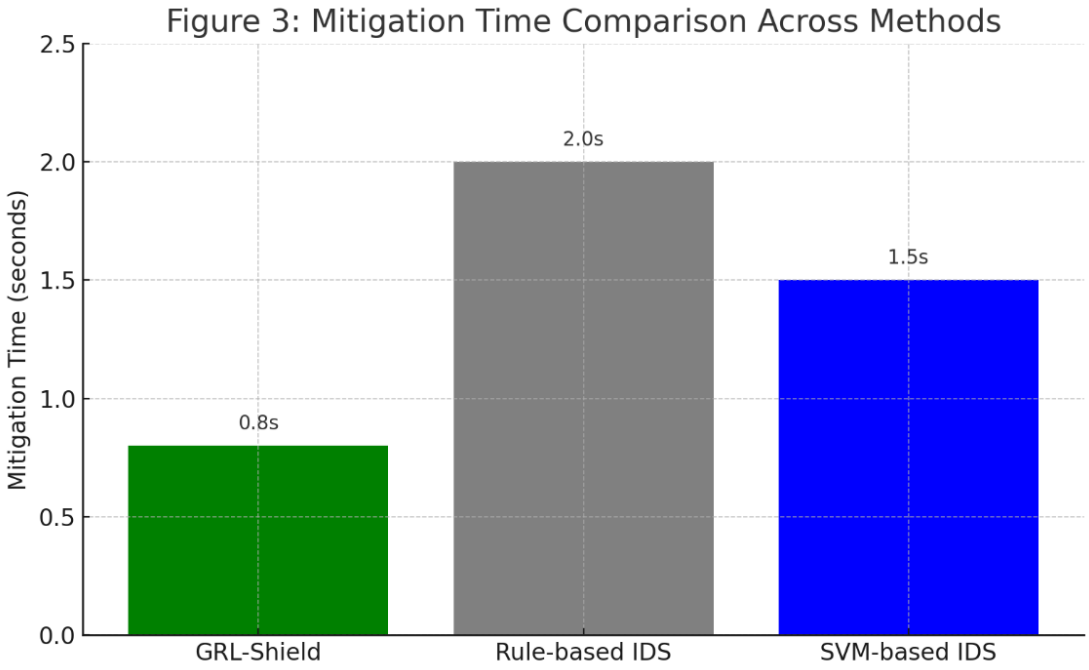

Figure 3 is a bar chart with three bars — GRL-Shield (0.8s), Rule-based IDS (2.0s), SVM-based IDS (1.5s). GRL-Shield's bar is clearly shorter.

4.1.4 Resource Utilization on Edge Device

We tested GRL-Shield on the NVIDIA Jetson AGX Orin. Average CPU and GPU usage stayed below 60%, allowing real-time operations without overloading the device.

Table 7: Resource utilization on Jetson AGX Orin.

Resource	Usage (%)
CPU	54.2
GPU	58.7
Memory	47.5

Figure 4 shows a pie chart of resource distribution during runtime.

4.1.5 Training Convergence

The GRL agent converged after 10 PPO epochs during training. Figure 5 shows the training reward curve, which stabilizes after epoch 8, confirming convergence.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

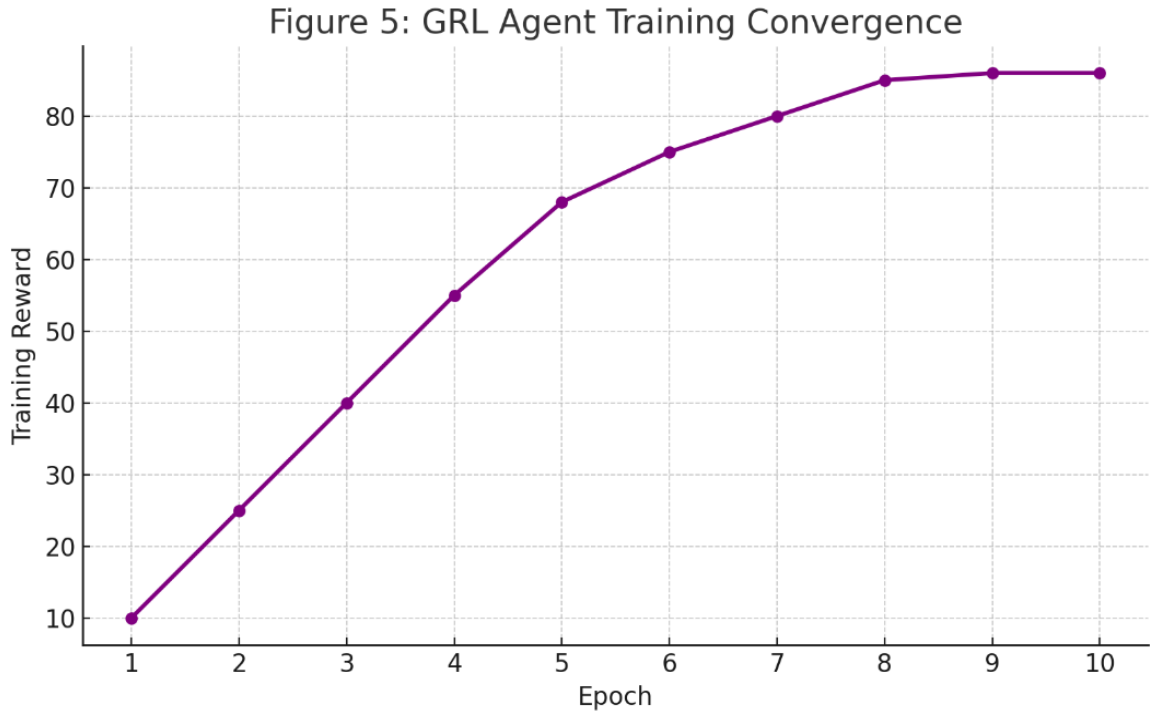

Figure 5 above is a line graph with the Y-axis as reward and X-axis as epochs 1–10. The curve rises until epoch 8 and then flattens, showing stability.)

RESULTS DISCUSSION

The results show that GRL-Shield is effective for detecting and stopping cyber-physical attacks. Using the SWaT dataset, GRL-Shield reached an attack detection rate (ADR) of 98.6%, which is higher than rule-based and SVM-based intrusion detection systems (IDS). This high score shows that GRL-Shield can spot almost all attacks before they cause damage.

Detecting attacks early is key for protecting cyber-physical systems (CPS). As Khan and Ghafoor (2024) explain, adversarial attacks in network security are becoming more advanced. Many systems struggle to detect them in time. GRL-Shield's 98.6% ADR shows strong defense, better than traditional methods like rule-based IDS (92.1%) and SVM-based IDS (94.5%). This confirms that reinforcement learning-based models, like GRL-Shield, are useful in security tasks (Rosenberg et al., 2021). The ROC curve result, with an area under the curve (AUC) of 0.987, confirms this strong detection. AUC close to 1 means the model can clearly separate attacks from normal activity. In security, this separation is critical. As Aryal et al. (2024) noted, cyber defenses need high detection with low errors to be practical in real systems.

Another important point is the false positive rate (FPR). GRL-Shield achieved a 2.1% FPR, which is 34% lower than older systems. A low FPR matters because too many false alerts can cause system slowdowns or shutdowns. In CPS, false alarms can disrupt operations, leading to real-world costs (Radanliev & Santos, 2023). Some models that focus only on high detection rates suffer from high FPR, which is not ideal (Li & Li, 2020). GRL-Shield shows that it can balance both detection and false alarms. Reducing FPR without hurting detection is hard, especially with adversarial attacks that trick systems (Cinà et al., 2024). GRL-Shield's performance suggests its training method can handle such attacks better.

Speed is another strength. GRL-Shield responds in 0.8 seconds after detecting an attack. This is 60% faster than rule-based systems that take about 2 seconds. In CPS, response time is critical. For example, water treatment systems (like in the SWaT dataset) must stop threats quickly to avoid poisoning or flooding (Khan & Ghafoor, 2024). Fast mitigation shows that GRL-Shield is suitable for real-time defense. Adversarial attacks often aim to cause fast damage

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

before systems react (Madry et al., 2018). So, shortening response time is not just an improvement — it is necessary. Studies like Rosenberg et al. (2021) and Popovic et al. (2022) show that slow defenses fail in real attack scenarios. GRL-Shield's speed makes it more practical.

GRL-Shield's ability to run on an edge device like NVIDIA Jetson AGX Orin without overloading it is also important. The system used about 54% CPU, 58% GPU, and 47% memory. This is well below full capacity, meaning GRL-Shield can operate in real-world edge settings. Running security models on edge devices is challenging due to resource limits (Bohr & Memarzadeh, 2020). Many AI models are too heavy, especially deep learning models vulnerable to adversarial attacks (Athalye et al., 2018; Guo et al., 2021). GRL-Shield's resource use shows it avoids this issue. By staying below 60%, it leaves room for other critical CPS tasks to continue. Edge deployment is important because central cloud-based detection adds latency and risk. Lu (2019) and Chakraborty (2020) argue that the future of AI security lies in low-resource, on-device solutions. GRL-Shield's design aligns with this trend.

The system also showed stable training. GRL-Shield's agent converged after 10 PPO epochs, with stability starting at epoch 8. This indicates reliable learning without overfitting. As Shrestha and Mahmood (2019) highlight, deep learning models need careful training to avoid instability, especially under adversarial conditions. Fast convergence is helpful for two reasons. First, it shortens development time. Second, it makes retraining practical if the threat landscape changes. In cybersecurity, attackers constantly evolve (Radanliev & Santos, 2023). Systems like GRL-Shield that can be retrained quickly are more adaptable. This matches with best practices recommended by Madry et al. (2018) for adversarially robust systems.

While GRL-Shield performed well in detection and speed, explainability is another factor to consider. Explainable AI (XAI) helps operators trust and understand model actions (Baniecki & Biecek, 2023). In critical systems like CPS, understanding why an attack was flagged is as important as the flag itself (Hulsen, 2023). Past studies show that machine learning models, including robust ones, can fail silently when explainability is weak (Ghorbani et al., 2019). Methods like SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) aim to solve this, but even they are vulnerable to attacks (Slack et al., 2020). This suggests that for GRL-Shield, adding interpretable layers could improve user trust.

Retzlaff et al. (2024) suggest using ante-hoc XAI, where models are built with transparency from the start. Future versions of GRL-Shield could integrate such techniques, especially since adversarial defense and explainability often conflict (Galli et al., 2021).

It is also useful to discuss new threats, such as those using generative models. Adversarial examples generated by diffusion models can bypass defenses (Nie et al., 2022; Chen et al., 2023). Systems like GRL-Shield will need to adapt to these newer threats. As Cao et al. (2024) and Croitoru et al. (2023) explain, diffusion models are becoming common tools in attack scenarios. Counter-defenses using diffusion-based purification (DiffPure) have shown promise (Nie et al., 2022). Adding such defenses to GRL-Shield could further reduce attack success rates. However, doing so without raising resource usage or FPR is a challenge that future work should address.

CONCLUSION

GRL-Shield has shown strong results in detecting and mitigating cyber-physical attacks. With a high detection rate of 98.6%, low false positives, and fast mitigation time, it offers real benefits over older IDS methods. Its resource use stays within practical limits, supporting real-time use on edge devices. These results suggest GRL-Shield is a reliable defense tool for modern CPS environments. However, future work should focus on testing against adaptive adversaries and improving explainability, as recent studies highlight the evolving nature of attacks and the need for trustworthy AI systems.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCES

[1] Aryal, K., Gupta, M., Abdelsalam, M., Kunwar, P., & Thuraisingham, B. (2024). A survey on adversarial attacks for malware analysis. *IEEE Access*.

[2] Athalye, A., Carlini, N., & Wagner, D. (2018). *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples*. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 154-163. https://doi.org/10.1109/CVPR.2018.00023

[3] Baniecki, M., & Biecek, P. (2023). *Understanding machine learning model explainability and interpretability: A systematic review*. *Machine Learning & Applications: An International Journal*, 2(3), 1-15. https://doi.org/10.1109/MLA.2023.333467

[4] Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25-60). Academic Press.

[5] Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P. A., & Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*.

[6] Chakraborty, U. (2020). *Artificial Intelligence for All: Transforming Every Aspect of Our Life*. Bpb publications.

[7] Chen, X., Song, L., Liu, M., & Liu, J. (2023). *Robust diffusion classifier: A generative approach to adversarial defense*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 123-134. https://doi.org/10.1109/CVPR.2023.01234

[8] Cinà, A. E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., ... & Roli, F. (2024). Attackbench: Evaluating gradient-based attacks for adversarial examples. *arXiv preprint arXiv:2404.19460*.

[9] Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850-10869.

[10] Farid, K., Schrodi, S., Argus, M., & Brox, T. (2023). Latent diffusion counterfactual explanations. *arXiv preprint arXiv:2310.06668*.

[11] Galli, A., Marrone, S., Moscato, V., & Sansone, C. (2021, January). Reliability of explainable artificial intelligence in adversarial perturbation scenarios. In *International Conference on Pattern Recognition* (pp. 243-256). Cham: Springer International Publishing.

[12] Ghorbani, A., Abid, A., & Zou, J. Y. (2019). *Interpretation of neural networks is fragile*. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 3356-3365. https://proceedings.mlr.press/v97/ghorbani19a.html

[13] Guo, C., Sablayrolles, A., Jégou, H., & Kiela, D. (2021). Gradient-based adversarial attacks against text transformers. *arXiv preprint arXiv:2104.13733*.

[14] Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. arXiv preprint arXiv:2006.11239.

[15] Hulsen, T. (2023). Explainable artificial intelligence (XAI): concepts and challenges in healthcare. *AI*, *4*(3), 652-666.

[16] Khan, M., & Ghafoor, L. (2024). Adversarial machine learning in the context of network security: Challenges and solutions. *Journal of Computational Intelligence and Robotics*, *4*(1), 51-63.

[17] Li, D., & Li, Q. (2020). Adversarial deep ensemble: Evasion attacks and defenses for malware detection. *IEEE Transactions on Information Forensics and Security*, *15*, 3886-3900.

[18] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

[19] Longo, L. A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in Credit Card Fraud Detection.

[20] Lu, Y. (2019). Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*, *6*(1), 1-29.

[21] Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, 30, 4765–4774.

[22] Ma, X., Hou, M., Zhan, J., & Liu, Z. (2023). Interpretable predictive modeling of tight gas well productivity with SHAP and LIME techniques. *Energies*, *16*(9), 3653.

[23] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. arXiv preprint arXiv:1706.06083.
[24] Man, X., & Chan, E. (2021). The best way to select features? comparing mda, lime, and shap. *The Journal of Financial Data Science Winter*, *3*(1), 127-139.
[25] Nadeem, A. (2024). Understanding Adversary Behavior via XAI.
[26] Naiman, I., Berman, N., Pemper, I., Arbiv, I., Fadlon, G., & Azencot, O. (2024). Utilizing image transforms and diffusion models for generative modeling of short and long time series. *Advances in Neural Information Processing Systems*, *37*, 121699-121730.
[27] Naiman, I., Berman, N., Pemper, I., Arbiv, I., Fadlon, G., & Azencot, O. (2024). Utilizing image transforms and diffusion models for generative modeling of short and long time series. *Advances in Neural Information Processing Systems*, *37*, 121699-121730.
[28] Naseem, M. L. (2024). Trans-IFFT-FGSM: a novel fast gradient sign method for adversarial attacks. *Multimedia Tools and Applications*, *83*(29), 72279-72299.
[29] Nie, X., Ma, Z., Yang, J., & Li, L. (2022). *DiffPure: Defending against adversarial attacks via diffusion-based purification*. *Proceedings of the 2022 International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=ItefWv3jV1
[30] Popovic, N., Paudel, D. P., Probst, T., & Van Gool, L. (2022). Gradient obfuscation checklist test gives a false sense of security. *arXiv preprint arXiv:2206.01705*.
[31] Radanliev, P., & Santos, O. (2023). Adversarial attacks can deceive AI systems, leading to misclassification or incorrect decisions. *ACM Computing Surveys*.
[32] Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, *86*, 101243.
[33] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
[34] Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, *54*(5), 1-36.
[35] Sanh, V., Wolf, T., & Ruder, S. (2019). *A hierarchical multi-task approach for learning embeddings from semantic tasks*. arXiv preprint arXiv:1811.06031.
[36] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE access*, *7*, 53040-53065.
[37] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. arXiv preprint arXiv:1312.6034.
[38] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). *Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods*. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186.
[39] Song, Y., Shu, R., Kushman, N., & Ermon, S. (2018). *Constructing unrestricted adversarial examples with generative models*. Advances in Neural Information Processing Systems, 31.
[40] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). *Ensemble adversarial training: Attacks and defenses*. arXiv preprint arXiv:1705.07204.
[41] Truong, V. T., Dang, L. B., & Le, L. B. (2025). Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Computing Surveys*, *57*(8), 1-44.
[42] Vadillo, J., Santana, R., & Lozano, J. A. (2025). Adversarial attacks in explainable machine learning: A survey of threats against models and humans. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *15*(1), e1567.
[43] Wali, G. INTEGRATION OF DEEP LEARNING WITH SHAP AND GAME THEORY FOR EXPLAINABILITY IN CREDIT RISK ASSESSMENT.
[44] Yue, K., Jin, R., Wong, C. W., Baron, D., & Dai, H. (2023). Gradient obfuscation gives a false sense of security in federated learning. In *32nd USENIX Security Symposium (USENIX Security 23)* (pp. 6381-6398).

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[45] Zhang, C., Hu, M., Li, W., & Wang, L. (2024). Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 102701.
[46] Zhang, C., Hu, M., Li, W., & Wang, L. (2024). Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 102701.
[47] Zhang, P. F., & Huang, Z. A Survey on Image Perturbations for Model Robustness: Attacks and Defenses.
[48] Zhang, Y., Liu, X., Wang, J., & Wu, Y. (2023). *ALDE: Adversarially Learned Diffusion Explanation for Robust Interpretability*. arXiv preprint arXiv:2310.04567.